TimeExpert: An Expert-Guided Video LLM for Video Temporal Grounding

Supplementary Material

Outline

In this Supplementary Material, we first provide details regarding the training algorithm of *TimeExpert* in Section 1. After that, we elaborate on the data preparation process in Section 2. We also present more details regarding our implementation and more experimental results in Section 3.

1. Detailed Training Algorithm of TimeExpert

Algorithm 1 Pseudo code of TimeExpert's MoE training process.

Require: Training data $\mathcal{D}$, gating network parameters $\mathbf{W}_g$, expert activation thresholds $\mathbf{G}$, task activation tracker A_t, learning rate η.
Ensure: Optimized MoE model parameters.

for each training iteration **do**
  for each mini-batch $(\mathbf{x}, y) \sim \mathcal{D}$ **do**
    Compute gating scores: $s(\mathbf{x}) = \frac{\langle \mathbf{x}, \mathbf{W}_g \rangle}{\|\mathbf{x}\| \|\mathbf{W}_g\|}$
    Apply task-aware gating:
    $g(\mathbf{x}) = \text{sign}\left(\sigma\left(\frac{s(\mathbf{x}) + \alpha A_t}{1+\alpha}\right) - \sigma(\mathbf{G})\right)$
    Select active experts: $\mathcal{E} = \{e \mid g(\mathbf{x})_e > 0\}$
    if $\mathcal{E} = \emptyset$ **then**
      Assign $\mathbf{x}$ to least-utilized expert e_{low} and update $A_{e_{\text{low}}}$
    end if
    Compute final MoE output: $\mathbf{y} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} g(\mathbf{x})_e \mathbf{E}_e(\mathbf{x})$
    Compute task-dependent auxiliary loss:
    $\mathcal{L}_{aux} = \lambda_1 \sum_{e=1}^{K} \left(\frac{A_e}{\sum_{j=1}^{K} A_j} - \frac{N_e}{\sum_{j=1}^{K} N_j} \right)^2 + \lambda_2 \sum_{e=1}^{K} ||W_{g,e}||_2$
    Compute total loss: $\mathcal{L} = \mathcal{L}_z$[24] $+ \mathcal{L}_{aux}$
    Update model parameters: $\mathbf{W}_g \leftarrow \mathbf{W}_g - \eta \nabla_{\mathbf{W}_g} \mathcal{L}$
    Update activation tracker: $A_e \leftarrow A_e + \sum_{\mathbf{x} \in \mathcal{E}} 1$
  end for
  if Iteration $\mod T_{\text{update}} = 0$ **then**
    Remove inactive experts: $\mathcal{E}_{\text{remove}} = \{e \mid A_e < \tau_{\text{min}}\}$
    Introduce new experts for unassigned task tokens
  end if
end for

2. Data Preparation

Stage 1: Task Module Pretraining. This stage focuses on equipping the model with fundamental multi-modal alignment and video-language capabilities. We employ datasets such as Valley [10], LLaVA_Image [23], TextVR [20], ShareGPT4Video [4], and VTG-IT [6] to establish a robust visual-text grounding foundation.

Stage 2: MoE Pretraining. To enable dynamic expert specialization, we introduce an MoE-based training phase leveraging diverse video-language datasets. Four subsets of Valley, TextVR, ShareGPT4Video, and VTG-IT are used alongside additional datasets such as ActivityNet Captions [2], VideoChatGPT [11], InternVid [18], and Next-QA [21]. These datasets provide extensive video-text interactions, enhancing the model's structured event reasoning capabilities.

Stage 3: Supervised Fine-tuning. The final training stage refines the model for downstream fine-grained video understanding. We incorporate datasets such as Moment-10M [14], EgoQA [12], STAR [19], and LLaVA-Video-178K [23]. This stage fine-tunes the model for precise event localization and dense video captioning.

2.1. Data Format

The annotations can be categorized into the following four types:

(1) **General Video Understanding.** General tasks include video captioning, image captioning, and video question answering, where the answer component does not contain timestamps or scores. Following [7], we employ a single token ⟨sync⟩ as a placeholder for these missing values, indicating an empty response. Datasets in this category include LLaVA_Image [23], Valley [10], TextVR [20], ShareGPT4Video [4], VideoChatGPT [11], and Next-QA [21].

(2) **Dense Video Captioning.** This task consists of timestamps and textual captions, where the ⟨sync⟩ token is used as a placeholder for scores. Relevant datasets include $\text{HiREST}_{\text{step}}$ [22], COIN [16], ActivityNet Captions [2], VTG-IT-DVC [6], and InternVid [18].

(3) **Moment Retrieval.** Similar to Dense Video Captioning, Moment Retrieval comprises timestamps and textual captions. This task includes $\text{HiREST}_{\text{grounding}}$ [22], QuerYD [13], DiDeMo [1], VTG-IT-MR [6], and InternVid [18].

(4) **Video Highlight Detection.** In this task, highlight moments are retrieved using textual queries. We employ the VTG-IT-VHD [6] dataset for this purpose.

2.2. Data Processing

To ensure high-quality and efficient training for TimeExpert, we leverage a structured data processing pipeline to balance data diversity while maintaining task relevance. This is crucial for video-language tasks requiring fine-grained temporal reasoning.

Given the varying reliability of existing video-language datasets, we employ automatic filtering to remove noisy or low-quality samples: (1) *Temporal Alignment Check:* We discard annotations with misaligned timestamps, ensuring accurate event descriptions match the corresponding video segments. (2) *Instruction Consistency Verification:* For datasets with multiple annotation sources, we use an LLM-based consistency check to filter ambiguous or overly generic captions. (3) *Scene Change Detection:* We apply optical flow analysis to eliminate videos with abrupt cuts that disrupt temporal coherence in event-based tasks.

It is worth noting that many datasets in Stage 3 do not have annotations in the style of VTG-IT. Therefore, we re-annotated a high-quality subset using mentioned datasets. Specifically, we employ Video-LLM-based iterative event grouping and extrapolative rule-based timestamp refinement to acquire precise video events and their corresponding timestamps.

This comprehensive data processing strategy ensures TimeExpert is trained on a well-curated, diverse, and high-quality dataset optimized for multi-task video-language understanding.

3. More Experiments

3.1. More Implementation Details

In line with prior work [6, 7], we standardize the representation of timestamps and scores to a fixed length format, consisting of four integer components, a decimal point, and one fractional component. To structure the sequence, we insert the token $\langle sep \rangle$ between consecutive timestamps or scores and append $\langle sync \rangle$ at the end of the sequence. For example, given the timestamp inputs $[9.56, 102.84]$, the corresponding tokenized sequence is:

$$\langle 0 \rangle\langle 0 \rangle\langle 9 \rangle\langle . \rangle\langle 5 \rangle\langle 6 \rangle\langle sep \rangle\langle 1 \rangle\langle 0 \rangle\langle 2 \rangle\langle . \rangle\langle 8 \rangle\langle 4 \rangle\langle sync \rangle.$$

Moreover, we incorporate temporal information by encoding each frame's timestamps with a time encoder. After discarding the ⟨sync⟩ and ⟨sep⟩ tokens, we obtain 6 time tokens per frame, which are concatenated with the 8 compressed visual tokens to form the final visual input. All experiments are conducted on 2 nodes of 8 NVIDIA H100-80GB GPUs. The hyperparameter settings are shown in Table 1.

3.2. More Experimental Results

Performance against Strong Generalist Video-LLMs. In this experiment, we select Qwen2-VL-Instruct [17] as a representative baseline of generalist Video-LLMs. ARIA [8] is also included since it serves as the foundation of our MoE decoder. We evaluate TimeExpert's temporal reasoning capability using two VTG-specific benchmarks, TempCompass [9] and TemporalBench [3], while also assessing its general video understanding a ability through general video question-answering benchmarks like VideoMME [5]. For VideoMME [5], our model matches ARIA's performance, indicating that specializing for time-sensitive tasks does not degrade, and even improve general video understanding abilities. For TempCompass [9], which assesses fine-grained temporal reasoning, TimeExpert improves the subtask score compared to ARIA, with notable gains in Action (+4.52) and Speed (+4.22), reinforcing its ability to localize dynamic events with precision. For TemporalBench [3], our model outperforms all baselines, particularly in Captioning, confirming its strong VTG-specific capabilities. These results demonstrate the effectiveness of our MoE-based structured modeling, which enhances both general video understanding and task-specific VTG performance.

Qualitative Comparisons. We compare the proposed TimeExpert with three state-of-the-art methods: TimeChat [15] and TRACE [7], which are VTG-specific Video-LLMs, as well as Qwen2-VL-7B-Instruct, a general-purpose Video-LLM. As illustrated in Figure 1, TimeExpert produces a highly precise prediction (62.05s to 65.98s), closely aligning with the ground truth. In contrast, Qwen2-VL, a general-purpose Video-LLM, fails to output structured timestamps and instead generates an open-ended textual response that does not directly provide a temporal grounding for the query. TRACE, identifies part of the relevant event but generates a significantly earlier starting timestamp (54.28s), leading to inaccurate localization. TimeChat, while correctly identifying the dosa batter as the relevant substance, misinterprets an earlier step in the preparation process as the target action, leading to an incorrect time span (24.78s to 44.70s).

References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 1

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 1

[3] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 2, 3

Model Configuration		Training Configuration		Dataset Configuration	
Hyperparameter	**Value**	**Hyperparameter**	**Value**	**Hyperparameter**	**Value**
Intermediate Size for MoE Layers	1664	Per-Device Train Batch Size	8		
No. of Experts in Each MoE Layer	64	Gradient Accumulation Steps	2		
No. of Shared Experts	2	Learning Rate	5×10^{-6}		
MoE Z-Loss Coefficient	1×10^{-5}	Weight Decay	0.1		
MoE Auxiliary Loss Coefficient	1×10^{-3}	Adam Beta2	0.95	No. of Dataloader Workers	8
Maximum Image Size	980	Warm-up Ratio	0.01		
Freeze ViT	True	LR Scheduler Type	Cosine		
Freeze LLM	False	Number of Training Epochs	1		
Freeze Projector	False	Maximum Sequence Length	2048		

Table 1. **Summary of hyperparameters.** This table includes **Left:** Model Configuration, **Middle:** Training Configuration, and **Right:** Dataset Configuration.

Method	Short	Medium	Long
ARIA [8]	**76.9**	67.0	58.8
Qwen2-VL-Instruct [17]	–	–	–
TimeExpert (Ours)	<u>76.8</u>	**67.1**	**58.9**

(a) VideoMME (w/o subs) [5]

Action	Direction	Speed
91.19	51.38	58.12
91.87	**54.90**	54.73
95.71	<u>54.50</u>	**62.34**

(b) TempCompass [9]

Short	Long	Captioning
26.6	23.5	51.5
24.7	18.8	51.9
26.6	**24.7**	**56.5**

(c) TemporalBench [3]

Table 2. **Quantitative Comparison of *TimeExpert* against Strong Generalist Video-LLMs across General Video Understanding Benchmarks and VTG-specific Video Benchmarks.** The best results are in **bold**. Some second-best results are marked with <u>underline</u>. For brevity, we display only a subset of TempCompass.

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Zhenyu Tang, Li Yuan, et al. Sharegpt4video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, pages 19472–19495, 2024. 1

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118, 2025. 2, 3

[6] Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI conference on artificial intelligence*, 2025. 1, 2

[7] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. Trace: Temporal grounding video llm via causal event modeling. In *International Conference on Learning Representations*, 2025. 1, 2, 4

[8] Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Guoyin Wang, Bei Chen, and Junnan Li. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 2, 3

[9] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of Association for Computational Linguistics*, 2024. 2, 3

[10] Ruipu Luo, Ziwang Zhao, Min Yang, Junwei Dong, Da Li, Pengcheng Lu, Tao Wang, Linmei Hu, Minghui Qiu, and Zhongyu Wei. Valley: Video assistant with large language model enhanced ability. *arXiv preprint arXiv:2306.07207*, 2023. 1

[11] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Association for Computational Linguistics*, 2024. 1

[12] Thong Thanh Nguyen, Zhiyuan Hu, Xiaobao Wu, Cong-Duy T Nguyen, See-Kiong Ng, and Anh Tuan Luu. Encoding and controlling global semantics for long-form video question answering. *arXiv preprint arXiv:2405.19723*, 2024. 1

[13] Andreea-Maria Oncescu, Joao F Henriques, Yang Liu, Andrew Zisserman, and Samuel Albanie. Queryd: A video dataset with high-quality text and audio narrations. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2265–2269. IEEE, 2021. 1

[14] Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, Yueting Zhuang, and Siliang Tang. Momentor: Advancing video large language model with fine-grained

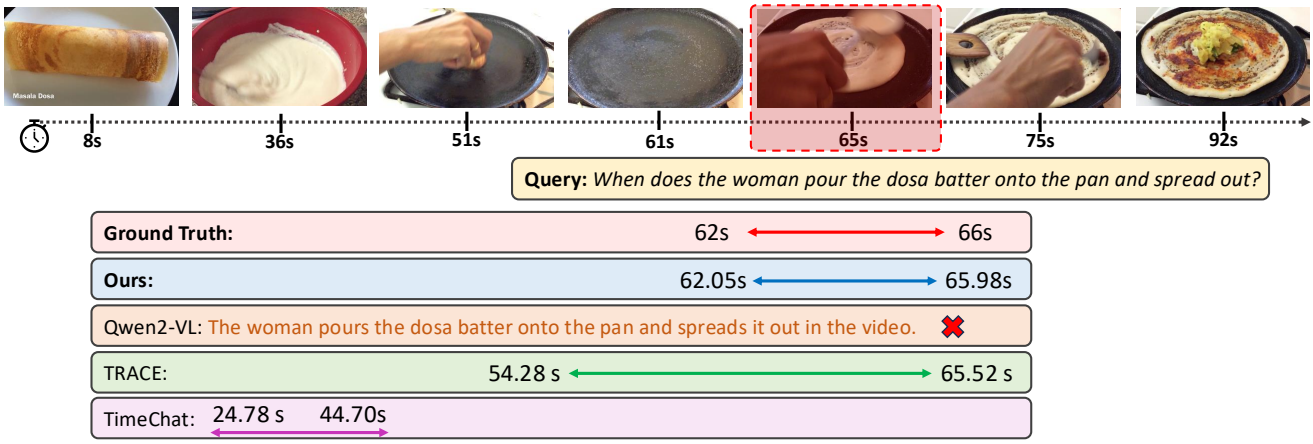

Figure 1. **Qualitative Comparison of *TimeExpert* against Several state-of-the-art Video-LLMs.** TimeExpert outperforms both general-purpose Video-LLMs (e.g., Qwen2-VL-Instruct [17]) and VTG-specific Video-LLMs (e.g., TimeChat [15], TRACE [7]) by precisely identifying event boundaries.

temporal reasoning. In *International Conference on Machine Learning*, 2024. 1

[15] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2, 4

[16] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 1

[17] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3, 4

[18] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 1

[19] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*, 2024. 1

[20] Weijia Wu, Yuzhong Zhao, Zhuang Li, Jiahong Li, Hong Zhou, Mike Zheng Shou, and Xiang Bai. A large cross-modal video retrieval dataset with reading comprehension. *Pattern Recognition*, 157:110818, 2025. 1

[21] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786, 2021. 1

[22] Abhay Zala, Jaemin Cho, Satwik Kottur, Xilun Chen, Barlas Oguz, Yashar Mehdad, and Mohit Bansal. Hierarchical video-moment retrieval and step-captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23056–23065, 2023. 1

[23] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1

[24] Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*, 2022. 1